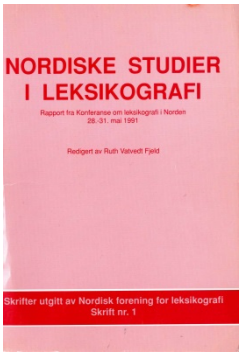
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: En ordbog er en database

Forfatter: Helle Degnbol, Guðrùn Ása Grímsdóttir, Bent Chr. Jacobsen, Jette Knudsen, Eva Rode & Christopher Sanders

Kilde: Nordiske Studier i Leksikografi 1, 1992, s. 385-389
Rapport fra Konferanse om leksikografi i Norden, 28.-31. mai 1991

URL: http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive



Søgbarhed
Artiklerne i de ældre Nordiske studier i leksikografi (1-5) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Helle Degnbol, Guðrún Ása Grímsdóttir, Bent Chr. Jacobsen, Jette Knudsen, Eva Rode & Christopher Sanders

En ordbog er en database

Under produktionen af *Ordbog over det norrøne prosasprog: Registre* (første bind af Den arnamagnæanske kommissions ordbog i 12 bind, udkommet i slutningen af 1989) lagdes alle oplysninger om ordbogens kildemateriale (håndskrifter, udgaver, norrøne og udenlandske) i database. Fra databaserne udvalgte redaktionen, hvilke data pr. automatik skulle bringes i registerbindet og under hvilken præsentationsform (typografi, layout). Det traditionelle stadium i en bogs tilblivelsesproces med udfærdigelse af "tekstbehandlet" manuskript og efterfølgende korrekturlæsning kunne således overspringes, og arbejdet kunne koncentreres om bogens indhold.

For halvandet år siden udkom det første af tolv ordbogsbind: *Ordbog over det norrøne prosasprog: Registre / A Dictionary of Old Norse Prose: Indices*. Bindet præsenterer ordbogens korpus, det norrøne (dvs. det middelalderlige islandsk-norske) prosasprog, i form af et signaturregister med oversigt over benyttede udgaver og håndskrifter (selve "nøglen"), et register over benyttede middelalderhåndskrifter med dateringer og litteraturhenvisninger, samt et register over benyttet litteratur, første del af ordbogens akkumulerende bibliografi.

For en kort beskrivelse af ordbogsprojektet henvises til *Lexikonord: Leksikografi i Norden*, Nordisk språksekretariats rapporter 14 (Oslo 1991:67-71).

Specielt om baggrunden for registerbindet og dets produktion kan man læse i The Arnamagnæan Institute & Dictionary: *Bulletin 17*: 1988-1989 (København 1990:23-27).

Prøver på ordbogsartikler ses i *Ordbog over det norrøne prosasprog: Prøvehæfte / A Dictionary of Old Norse Prose: Prospectus*, Kbh. 1983.

Al virksomhed er nu koncentreret om produktion af artikler til det næste ordbogsbind. I det følgende præsenteres glimtvis hvordan vi rent teknisk arbejder med at lave "egentlig" ordbog: i princippet som ved udarbejdelsen af registerbindet, dvs. i database fra første færd og så nær hen til den endelige bogproduktion som muligt, i et system, vi selv har opbygget og til stadighed udvikler.

Database

Hvad vi forstår ved en database, er en samling oplysninger indlagt i en fast struktur. Ordbogens database er emnemæssigt opdelt i enheder kaldet filer.

Centralt står ord-/hovedfilen med lemma, ordtypebestemmelse (til udskillelse af fx. person- og stednavne og poetisk ordforråd), grammatisk bestemmelse og oplysning om ordenes bøjning og formvariation (af denne fil dannes *artikelens hoved*), citatfilen og definitionsfilen (af hvilke dannes *artikelens krop*).

Perifert står filer med komposita-forled og med oplysninger om ordenes behandling i anden litteratur, herunder ordbøger (af hvilke filer dannes *artikelens hale*, afsnittene Comp., Litt. og Gloss.).

Ligeledes perifert står en fil med oplysning om de til grund for signaturerne liggende udgaver og håndskrifter, en værkfil, samt en håndskriftfil med oplysning om datering, proveniens mm. (Disse filer danner grundstammen i *registerbindet*.)

Endelig hører der til hele systemet en bibliografisk fil (efter hvilken den akkumulerende *bibliografi* trykkes).

Hver af disse filer er opdelt i mange, mindre felter, hvert felt rummende sin information. Således har alle oplysninger, store som små, deres plads i en ordnet struktur. Med sin konsistente arkitektur er systemet fleksibelt og giver mulighed for datavalidering på højt niveau, fx. kontrol af signaturer; udnyttelse af allerede lagrede informationer, fx. datering af citatmaterialet; beskuelse af materialet under forskellige synsvinkler; rettelser og omflytninger; vilkårligt bestemte udskrifter; automatisk registrering af hvert ords status i redaktionsprocessen. Alt bestemmes og udføres direkte af redaktørerne.

Citatmateriale

Kernen i ordbogsartikelen og dens hele udgangspunkt er citaterne, og den mest fundamentale enhed i databasen er den fil, der indeholder ordbogens citatmateriale.

ONP's seddelsamling rummer mange forskellige citattyper: enkle citater med kun én tekst (hentet enten i en udgave eller direkte i et håndskrift); citater med en eller flere varianter, repræsenterende forskellige versioner af samme tekst; citater med udenlandsk parallel; citater med udgivers rettelse; citater med udgivers kommentar; citater med ordbogsredaktørs kommentar, evt. i form af norrøn hjælpetekst eller i form af oversættelse til dansk/engelsk. Ofte er disse citattyper kombineret, og der er derfor i databasen brug for mange felter svarende til alle de forskellige lag og typer af oplysninger. Citatfilen har således henved 100 felter.

Databasen hjælper på forskellig vis til rationel behandling af materialet, og der skal her gives et enkelt eksempel (et mindre pilotprojekt): En type sedler, der traditionelt vil koste en del arbejde, er dem, der rummer citater afskrevet direkte fra håndskrifterne, idet de pågældende tekster endnu ikke er udgivet. Ordsedlerne indtastes, som de optræder i skufferne, dvs. med citater fra mange forskellige håndskrifter i tilfældig rækkefølge. Herpå sorteres de efter håndskrift og håndskriftside. Man har nu et praktisk overblik over materialet og kan sammenligne citat og originaltekst håndskrift for håndskrift og side for side, man kan nemt indsætte linjenummer, som er det der oftest mangler, kontrollere ortografien, tilføje mere kontekst osv. På denne måde er det kun nødvendigt at konsultere det samme håndskrift én gang.

På samme måde kan databasen sikre rationel korrekturlæsning af ordbogens mange citater imod udgaverne.

Når redaktøren modtager citaterne med henblik på redigering, er de blevet indtastet af assistenter. I samme proces er signaturerne kontrolleret og revideret og visse systematiske fejl påvist og udryddet.

For såkaldt mindre ords vedkommende indtastes alle ordsedler. Inden for bogstavet a er der ca. 40.000 citater fordelt på ca. 4.000 ord; 83 % af ordene er belagt med mindre end 9 eksempler, og hele 96,5 % af ordene er på under 50 belæg. De indtastes rub og stub. Inden for de såkaldt større ord, de 3,5 % af ordene, der er belagt med mere end 50 (enkelte med

over 1000) eksempler, sorteres der fra inden indtastning og redigering. Selv om det er ærgerligt at måtte undvære en del af citaterne i databasen, ses det klart, at frasorteringen betaler sig, da disse få større ord udgør hele 55 % af seddelmaterialet.

Citaterne kan redaktøren få i hænde (dvs. op på sin skærm eller printet ud i et hvilket som helst ønsket format, fx. seddelformat) sorteret efter eget valg: altid naturligvis efter ord (lemma), men dernæst fx. alfabetisk efter navnene på de værker, der er citeret (det kan være praktisk ved meget store ord) eller kronologisk efter de tilgrundliggende håndskrifters alder. Håndskriftdatering er automatisk indført i filen og påført udskriften, og ordets ældste citat er automatisk markeret, med et mærke, der følger citatet hele vejen ud i den trykte ordbog. Det vil altid være værdifuldt for redaktøren at få dette kronologiske overblik over ordet. Det kan også være frugtbart ved databasemæssig sortering at få samlet alle citater, til hvilke der er udenlandske paralleller, ligesom citaterne kan vælges sorteret efter værkgenre.

Redigering

Redigeringen er den mest udfordrende del af ordbogsproduktionen, både i sig selv og i forhold til databasesystemet.

Redigeringsarbejdet i alle dets faser foregår (eller kan foregå) i database. Mest aktivt redigeres der på de centrale filer: ord-/hoved-, definitions- og citatfilerne, som tilsammen skal udgøre artikelens hoved og krop.

Redaktøren kan i princippet begynde redigeringen hvorsomhelst i systemet, men han begynder gerne i *citatfilen*, ligesom man ved traditionel redigering oftest tager fat på citatmaterialet først og sorterer sedler. Ved hjælp af særlige redigeringsfelter foretager redaktøren den første analyse af citaterne og tager for hvert citats vedkommende stilling til ordets/citatets betydning og brug, indsætter evt. kommentar til eller særoversættelse i citatet, ser på ordets syntaktiske stilling og eventuelle indgåen i en ordforbindelse af løsere eller fastere karakter, samt danner sig et overblik over norrøne varianter og udenlandske paralleller.

Det er på dette stadium naturligt at formulere definitioner og beslutte sig for artikelens inddeling og opstilling, kort sagt lægge data i *definitionsfilen*. I citatfilens særlige indekseringsfelter markeres det, hvilke citater hører til hvilken betydningsafdeling (evt. syntaktisk afdeling eller afdeling med ordforbindelse), hvilke belæg skal citeres og i hvilken rækkefølge, hvilke blot bringes som 'nøgen henvisning' (signatur uden citat), og hvilke ikke skal med overhovedet. I det integrerede databasesystem ruller definitioner og citater koordineret hen over skærmen. Indekseringsfelterne gør det let og smidigt at ændre i artikelen, idet den så at sige sorterer sig selv. Fx. vil et citat ved ændring af dets betydningsnummer automatisk flytte plads, og det er tilsvarende nemt at ordne citaternes rækkefølge inden for de forskellige afdelinger. I hver afdeling følger efter citaterne et udvalg af nøgne henvisninger i kronologisk orden.

I databasesystemet får man et klart overblik over ordets ortografiske variation og dets bøjningsformer og kan dermed færdigudfylde *ord-/hovedfilen*.

Når centralfilerne er færdigbearbejdet, er artikelens hoved og krop formet. De afsnit, der udgør artikelens hale med stof fra både de centrale og de perifære filer, genereres næsten automatisk, dvs. med meget lidt manuel indblanding fra redaktørhold. Latinske, franske, tyske, engelske og danske paralleller indhentes fra citatfilen, sammensætningsforled fra kompositafilen og ordbøger/glossarer, der behandler det pågældende ord, fra glossarfilen. Hermed er artikelen klar til at forlade førsteredaktørens bord.

Redaktionelt samarbejde

Databasen byder på særlige muligheder, når flere redaktører vil forsøge sig med redigering af det samme intakte citatmateriale. Når førsteredaktøren af en artikel har lavet sit udkast, gennemgås det grundigt af andenredaktøren. I databasen kan andenredaktøren komme med egne forslag til beskrivelse af ordet og til strukturering og anden tilskæring af artikelen, og han kan med få tryk på tastaturet ved hjælp af egne indekseringsfelter afprøve sine ideer. Oftest, i de mange mindre artikler, viser han dog sine forslag blot ved at gøre antegninger i en arbejdsudskrift.

Arbejdsudskriften kan bestå ene og alene af artikelens skelet, dvs. dens afdelinger og definitioner; eller den kan have den længst mulige form, vise alle citater, også dem, der foreslås bragt som nøgne henvisninger, og dem, der foreslås udeladt, samt alle interne kommentarer; eller udskriften kan være udformet præcis som en artikel i den trykte bog.

Samme mulighed som andenredaktøren har den engelske redaktør, der gennemser alle artikler, først og fremmest med henblik på de danske/engelske definitioner. Den engelske redaktør kan vælge at arbejde i et særligt skærmformat, hvor engelsk tekst praktisk kan indsættes og kommentarer tilføjes. Undertiden gives der danske/engelske kommentarer eller oversættelser nede i selve citaterne, også til det formål er der udviklet skærmformater, og et lille program finder vej til de få steder, det drejer sig om.

Fra database til trykt bog

Takket være databasesystemet kan ordbogens medarbejdere indsamle og indskrive materiale, automatisk indhente oplysninger fra tilkoblede registre, analysere på materialet, redigere det, samle baggrundsoplysninger, diskutere med hinanden, gøre interne notater, alt sammen på en særdeles praktisk måde.

Arbejdet foregår inden for et par simple business-databasesystemer: dBASE III og SmartWareII. Vi vil, på vej til fotosætningen, benytte tekstformatteringssystemet TEX, som også blev benyttet ved fremstillingen af registerbindet (interface: Postscript). Helt endelig beslutning om typografi behøver man ikke træffe før sent i forløbet. Hvis bare man har stoffet fornuftigt struktureret og fordelt på diskrete kategorier, kan alt lade sig gøre typografisk. Tegnproblemer og alfabetiseringsproblemer er løst. Systemet er helt åbent. Det kan endnu ikke siges at fungere smidigt, men det vil blive smidigere efterhånden, og snart har vi, hvis vi skulle ønske det, et godt grundlag for at kunne definere et evt. skræddersyet ordbogssystem. Det, der ligger fast, er, at vi hele tiden grundigt kender og kan overskue og håndtere både system og materiale, og at vi hele tiden arbejder henimod nøjagtig det samme produkt. Som nævnt kan vi trække alskens arbejdsudskrifter ud af databasen undervejs, og vi kan definere et hvilket som helst skærmformat, men vigtigst af alt kan vi til sidst direkte ud af én og samme base trække det destillerede produkt: ordbogsartikelen og dermed ordbogen.

Som vi fik det afprøvet under arbejdet med registerbindet, hvor vi gik så direkte som overhovedet muligt fra database til trykt bog, indebærer en sådan bogproduktion indlysende praktiske fordele: ingen ventetider og mellemstationer, ingen traditionel tekstbehandling (det vil fx. sige ingen tegnsætning mellem de forskellige slags oplysninger) og ingen korrekturlæsning i traditionel forstand. Redaktørerne kan i stedet og indtil sidste øjeblik koncentrere sig om den vigtigere indholdskorrektur i selve databasen.

Vi har bevidst udskudt at arbejde med de allermest komplicerede artikeltyper, det gælder verberne med deres partikler og konstruktionseksempler. Men systemet er gjort klar til modtagelse og manipulering af også komplekse strukturer, og ikke mindst, når vi skal til

at håndtere disse mere tunge sager, har vi gode forhåbninger til databasesystemet både som opbevaringskasse og som redigeringsværktøj.

Når vi bliver spurgt, hvad vi vil stille op med de helt store ord, fx. de store præpositioner på over 1000 sedler, om vores databasesystem virkelig kan håndtere dem, så må vores svar være: I krig og kærlighed gælder alle kneb, og stillet over for særlig store leksikografiske problemer bør overhovedet ingen metode foragtes. Problemerne skal søges løst og artikelen præsenteres på den bedst mulige måde. Men grundtanken må fastholdes. Ikke nok med at databasesystemet er et fint stykke værktøj for mangen en videnskabsmand. En ordbog er en database, og i bund og grund betaler det sig at lade den være det fra først til sidst.

Litteratur

Nordisk språksekretariats rapporter 14. 1991. *Lexikonord.* Oslo
Ordbog over det norrøne prosasprog. 1983. *Prøvehæfte.* København
Ordbog over det norrøne prosasprog. 1989. *Registre.* København
The Arnamagnæan Institute & Dictionary. 1990. *Bulletin 17.* København